MycoKeys 53: 73–91 (2019)
doi: 10.3897/mycokeys.53.34761
http://mycokeys.pensoft.net

RESEARCH ARTICLE

A peer-reviewed open-access journal
MycoKeys
Launched to accelerate biodiversity research

PacBio amplicon sequencing for metabarcoding of mixed DNA samples from lichen herbarium specimens

Cécile Gueidan[1], John A. Elix[2], Patrick M. McCarthy[3], Claude Roux[4], Max Mallen-Cooper[5], Gintaras Kantvilas[3]

1 *Australian National Herbarium, National Research Collections Australia, CSIRO-NCMI, Canberra, ACT, 2601, Australia* **2** *Research School of Chemistry, Building 137, Australian National University, Canberra, ACT, 2601, Australia* **3** *64 Broadsmith St, Scullin, ACT, 2614, Australia* **4** *390 chemin des Vignes vieilles, 84120 Mirabeau, France* **5** *Centre for Ecosystem Science, School of Biological, Earth and Environmental Sciences, University of New South Wales Sydney, Kensington, NSW, 2052, Australia* **6** *Tasmanian Herbarium, Tasmanian Museum and Art Gallery, Sandy Bay, Tasmania 7005, Australia*

Corresponding author: *Cécile Gueidan* (Cecile.Gueidan@csiro.au)

Academic editor: *F. Dal Grande* | Received 22 March 2019 | Accepted 10 May 2019 | Published 3 June 2019

Citation: Gueidan C, Elix JA, McCarthy PM, Roux C, Mallen-Cooper M, Kantvilas G (2019) PacBio amplicon sequencing for metabarcoding of mixed DNA samples from lichen herbarium specimens. MycoKeys 53: 73–91. https://doi.org/10.3897/mycokeys.53.34761

Abstract

The detection and identification of species of fungi in the environment using molecular methods heavily depends on reliable reference sequence databases. However, these databases are largely incomplete in terms of taxon coverage, and a significant effort is required from herbaria and living fungal collections for the mass-barcoding of well-identified and well-curated fungal specimens or strains. Here, a PacBio amplicon sequencing approach is applied to recent lichen herbarium specimens for the sequencing of the fungal ITS barcode, allowing a higher throughput sample processing than Sanger sequencing, which often required the use of cloning. Out of 96 multiplexed samples, a full-length ITS sequence of the target lichenised fungal species was recovered for 85 specimens. In addition, sequences obtained for co-amplified fungi gave an interesting insight into the diversity of endolichenic fungi. Challenges encountered at both the laboratory and bioinformatic stages are discussed, and cost and quality are compared with Sanger sequencing. With increasing data output and reducing sequencing cost, PacBio amplicon sequencing is seen as a promising approach for the generation of reference sequences for lichenised fungi as well as the characterisation of lichen-associated fungal communities.

Keywords

SMRT sequencing, high-throughput sequencing, long amplicon analysis (LAA), lichenised fungi

Copyright *Cécile Gueidan et al.* This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Introduction

Across the world, herbaria and fungal culture collections host large numbers of specimens and strains with the main goal being to document, preserve and classify the many species within the fungal kingdom. Among characters generally used to identify fungi, DNA sequences have been particularly useful for the numerous fungal species with plastic or convergent morphologies. Of the DNA markers traditionally used for fungal identification, the internal transcribed spacer region (ITS) was chosen as the primary barcode because it allowed species-level identification for the broadest range of fungi (Schoch et al. 2012). Following the formal acceptance of this fungal barcode, substantial effort was put into improving the curation of ITS sequences in several databases, including RefSeq (Schoch et al. 2014; O'Leary et al. 2015), UNITE (Kõljalg et al. 2013; Nilsson et al. 2018) and ISHAM-ITS (Irinyi et al. 2015), by selecting high-quality ITS sequences generated from reliably identified and preserved specimens. These reference databases play a critical role in the sequence-based species identification of herbarium specimens, cultured strains and fungal communities from environmental samples (Tedersoo and Nilsson 2016). However, linking sequences to names remains a challenge, partly because of the incompleteness of these reference datasets (Orock et al. 2012; Kõljalg et al. 2013; Crous et al. 2014; Nilsson et al. 2018). The molecular barcoding of herbarium specimens, including generic and species types therefore remains a priority (Crous et al. 2014; Yahr et al. 2016).

Although the characterisation of fungal diversity in environmental samples has benefited immensely from the development of next generation sequencing (NGS) technologies, the high-throughput generation of reference ITS sequences from herbarium specimens has been hindered by the short length of Illumina reads, the most commonly used NGS platform. With a size ranging from 500 up to 1200 bp, the full length of the ITS region cannot be sequenced as a single DNA fragment. For DNA extractions of lichen specimens, which harbor on their surface or within their thalli a plethora of other fungi, as well as for any other DNA samples of mixed fungal communities, assembling DNA markers from a pool of short reads belonging to several species carries a high risk of obtaining chimeric sequences (Hebert et al. 2018). As a consequence, fungal metabarcoding studies mostly use only half of the ITS region, either ITS1 or ITS2 (Nilsson et al. 2010; Mello et al. 2011; Blaalid et al. 2013). The generation of full length and high-quality barcodes from lichen specimens for reference databases or identification purposes generally involved Sanger sequencing (e.g., Kelly et al. 2011; Orock et al. 2012; Leavitt et al. 2014; Divakar et al. 2016; Xu et al. 2017). However, this method faces a similar challenge: that co-amplified non-lichenised fungal sequences often prevent the generation of readable target sequences (Kelly et al. 2011; Orock et al. 2012). Of the few available methods to separate sequences from target and non-target species (e.g., gel separation, group-specific primers), cloning has proven to be the most broadly applicable method for obtaining high quality sequences of the target lichenised fungus in mixed samples (Hofstetter et al. 2007). It does, however, significantly increase the cost and time required for generating high-quality sequences.

Long-read sequencing allows us to circumvent these challenges. For lichenised fungi, early long amplicon sequencing studies made use of Roche 454 pyrosequencing technology, either for the full ITS region (Hodkinson and Lendemer 2013; Mark et al. 2016) or for ITS1 only (Lücking et al. 2014). In Mark et al. (2016), with a target sequence recovered for 99 of the 100 samples studied, the application of this technology for lichen specimen metabarcoding seemed promising. However, the development of this sequencing technology was abandoned by Roche and is now obsolete. Long read sequencing using single molecule real-time (SMRT) sequencing technology (Pacific Biosciences, USA) is becoming more affordable, and has recently been applied to fungi for metabarcoding purposes (Chen et al. 2015; Cline and Zak 2015; James et al. 2016; Schlaeppi et al. 2016; Walder et al. 2017; Heeger et al. 2018). Some of the challenges usually associated with SMRT sequencing (high error rate, high rate of chimeric formation and high cost per sample) have been investigated and partly addressed. The high error rate of SMRT raw reads (about 15%; Goodwin et al. 2016) is significantly decreased due to the multiple passes obtained from a single polymerase read using circularised amplifications (Travers et al. 2010), as well as the correcting nature of high sequence coverage for randomly distributed errors (Koren et al. 2012). The error rate of circular consensus sequences (CCS) is now usually under 1% (Goodwin et al. 2016), and compares well with other sequencing platforms, including Illumina (Schlaeppi et al. 2016; Schloss et al. 2016) and Sanger (Goodwin et al. 2016). Using mock communities, the rates of formation of chimeric sequences (up to 16.3% in Heeger et al. 2018) were shown to be in the same range as the one obtained with short read sequencing (D'Amore et al. 2016). As for the cost per sample, it can be reduced thanks to multiplexing (Heeger et al. 2018) to a level where the cost of DNA extraction or PCR amplification becomes higher than the sequencing cost (Hebert et al. 2018).

The scarcity of automated pipelines to analyse SMRT amplicon data for different applications, including community analysis, remains an issue (Heeger et al. 2018; Tedersoo et al. 2018). Most software developed for sequencing-error correction and sequence assembly have been optimised for short reads and have not been evaluated for SMRT raw data (Heeger et al. 2018). During the SMRT sequencing primary analysis, image processing, base-calling and quality assessment are done in real-time on the instrument. Polymerase reads are then used to generate CCSs, which, in addition to the raw polymerase read files, are provided by the sequencing facility as fasta or fastq files. For community analyses, CCSs can then be analysed by various software or software packages, such as Mothur (Schloss et al. 2009) or PipeCraft (Anslan et al. 2017), in order to demultiplex, filter, cluster and assign consensus sequences to a taxonomic unit (see Anslan et al. 2018). Although PacBio (Pacific Biosciences) provides secondary analysis modules as part of the SMRT portal, none of these is specifically designed for community analysis. The Long Amplicon Analysis (LAA) pipeline available on SMRT Link and SMRT Portal attracted our interest, as it identifies differing clusters of sequencing reads within a single library and is capable of differentiating between underlying sequences that are 99.9% similar, such as haplotypes and pseudogenes (Bowman et al. 2014). The main application of LAA is allele phasing and detection

in diploid organisms (Bowman et al. 2014; Lleras et al. 2014) and it works as follows. Error-corrected subreads (CCS) obtained from polymerases reads are first demultiplexed (samples are separated using unique molecular indexes or combinations of indexes), then filtered depending on read quality and length. They are then aligned (overlap step) and clustered based on the alignment. Each cluster is iteratively phased in order to separate high-scoring mutations (alleles). Resulting sub-clusters are polished using a Quiver-based method to produce high-quality consensus sequences, which then go through a last filtering step to detect and remove PCR artefacts (e.g., chimeric sequences). This pipeline was designed to differentiate sequencing errors from true sequence variations, allowing the preservation of all sequence variants within a sample. For each index/sample barcode, the output includes all unique consensus sequences obtained, together with their coverage value and predicted accuracies. LAA can be performed on the instrument's SMRT portal, allowing the sequencing provider to send these demultiplexed consensus sequences directly to the customer, saving time and effort in software installation and raw data analyses.

In this study, the use of SMRT sequencing and the LAA pipeline for the production of ITS barcode sequences from lichen herbarium specimens was explored. The goals were: 1) to establish a high-throughput protocol to obtain indexed PCR products from lichen DNA extracts for SMRT sequencing; and 2) to investigate the ability of LAA to recover high-quality sequences for the target lichenised fungal species as well as for non-target fungal species.

Material and methods

DNA extractions

A total of 96 specimens were selected for their frequent need for cloning as well as their relevance to several ongoing taxonomic studies on Australian lichens at the Australian National Herbarium (see Suppl. material 1: Table S1). The genera represented were *Catillaria* (41 specimens from Australia and 16 from France), *Buellia* (39 specimens from Australia), *Endocarpon* (8 specimens from Australia) and *Verrucaria* (2 specimens from France). The specimens were 1–34 years old and are kept at CANB, MARSSJ and in the private herbarium of M. Bertrand. For crustose specimens (*Catillaria*, *Buellia* and *Verrucaria*), material (thallus and fruiting bodies) was detached from the substrate with a clean single-edge razor blade and a folded sheet of weigh paper was used to collect and transfer the material to a tube (8-strip cluster tubes, Corning Incorporated, Salt Lake City, USA). For *Endocarpon*, squamules were detached from the soil substrate using clean tweezers and transferred directly to a tube.

Each of the 96 cluster tubes contained a washed chrome steel bearing ball 3 mm in diameter (3MMCH/S/B grade 40, BSC Bearing & Power Transmission Solutions, Chullora, Australia). The samples were ground with a TissueLyser II (Qiagen, Hilden, Germany) in two cycles of 1 min and a frequency of 25/s. The cluster tubes were

centrifuged for 1 min at 6,000 rpm and the caps were removed with care to avoid cross-contaminations. A lysis buffer was added to each tube. Genomic DNA was extracted using the Invisorb® DNA Plant HTS 96 kit (STRATEC Molecular, Berlin, Germany) following the manufacturer's instructions, except for the last centrifugation step which was changed to 10 min at 2,000 rpm (instead of 5 min at 4,000 rpm) to avoid breaking the elution plates. A 1/10 dilution of the DNA extraction plate was prepared and subsequently used for amplification.

Amplification, normalisation and pooling

Indexed PCR products were generated using the PacBio Barcoded Universal Primers protocol (https://www.pacb.com/wp-content/uploads/2015/09/Procedure-and-Checklist-Preparing-SMRTbell-Libraries-PacB-Barcoded-Universal-Primers.pdf). The ITS barcode (internal transcribed spacer 1, 5.8S ribosomal RNA subunit and internal transcribed spacer 2) was the region targeted. With a first PCR, our target region was amplified using the primers ITS1F (Gardes and Bruns 1993) and ITS4 (White et al. 1990), both modified by adding a 5' block and a tail representing the PacBio universal sequences. In a 25 µl reaction, 5 µl of buffer, 1 µl of MyFi (Bioline, London, UK), 1 µl of 5 µM of each primer, 16 µl of water and 1 µl of DNA template (1/10 dilution) were added. The 96 PCR reactions were performed in strip tubes with individual caps to avoid cross-contaminations. The PCR programme was 5 min at 95 °C, then 20 cycles of 30 sec at 95 °C, 30 sec at 53 °C and 1:30 min at 72 °C, followed by a final elongation step for 7 min at 72 °C. Four strips were randomly selected and their PCR products run on to an agarose gel using the nucleic acid stain GelRed (Biotium, Fremont, CA, USA). Because the gel showed primer dimer bands in addition to the amplicon bands, the PCR products were cleaned using Sera-Mag magnetic beads (SpeedBead Magnetic Carboxylate Modified Particles, GE Healthcare) and a 96 well Alpaqua® magnetic plate. The cleaning was done by adding 0.8× volume of beads to each PCR product, followed by two washes with 200 µl of 70% ethanol. Dry beads were then resuspended in 25 µl of the elution buffer from the Invisorb® DNA Plant HTS 96 kit.

A second amplification was then performed using the Barcoded Universal F/R Primers Plate-96 available from PacBio (Millenium Science, Mulgrave, VIC, Australia). In a 25 µl reaction, 5 µl of buffer, 1 µl of MyFi, 2.5 µl of the barcoded primers, 15.5 µl of water and 1 µl of the product of the first round of PCR were added. The PCR programme was 3 min at 95 °C, then 20 cycles of 30 sec at 95 °C, 30 sec at 53 °C and 1:30 min at 72 °C, followed by a final elongation step for 7 min at 72 °C. All PCR products were checked on a gel as previously described. For samples showing no band, a new round of amplification (first PCR with ITS tailed primers and second PCR with barcoded universal primers) was performed using the non-diluted DNA extracts. Positive PCR products generated by this second round of amplification were used to substitute the negative samples on the original plate. The samples for which neither the dilution nor the original extract gave amplicons were left on the original plate. The 96 samples

were cleaned as described above and their concentration measured with a Nanodrop 8000 spectrophotometer (Thermo Scientific, Waltham, MA, USA). All samples with a DNA concentration larger than 50 ng/µl were normalised manually to 50 ng/µl. One microliter of each of the 96 samples was pooled into a 1.5 ml Eppendorf tube.

Library preparation, sequencing and SMRT Portal primary analysis

The pooled sample (1.5 µg of DNA) was sent to the Ramaciotti Centre for Genomics (UNSW Sydney, Australia) for single molecular real-time (SMRT) sequencing. The sample met the quality control requirements and showed two clear peaks at 650 and 923 bp, which are within the expected size range for the ITS barcode. The library preparation was done using the PacBio Barcoded Universal Primers protocol (https://www.pacb.com/wp-content/uploads/2015/09/Procedure-and-Checklist-Preparing-SMRTbell-Libraries-PacB-Barcoded-Universal-Primers.pdf). The sample was sequenced in one SMRT cell on a PacBio RSII using a P6 chemistry with a four-hour movie. The raw data were analysed using Long Amplicon Analysis (LAA v. 1) with barcoding option on the SMRT Portal. This pipeline was run using the following settings: symmetric DNA barcodes with a minimum score of 22, a minimum subread length of 450 bp, a maximum number of subreads of 4,000, and default values for all other parameters. The phase alleles option was selected.

Molecular identification and validation

After demultiplexing, quality control and generation of consensus sequences using LAA, the amplicon data was provided in 96 files in fasta format. Sequences from the targeted lichenised fungal species were identified from fungal endophytes or contaminants based on sequence similarity using a BLASTN v 2.8.1 search on the nr database by excluding uncultured/environmental sample sequences (https://blast.ncbi.nlm.nih.gov/; search done on Dec 12–14 2018). The first BLASTN match (or highest bit score) was recorded unless it did not correspond to at least a fungal Class, in which case the next match was considered. For 12 samples, the target sequence obtained with PacBio was cross-validated by a sequence obtained with Sanger sequencing. The ITS barcode was amplified with the primers ITS1F and ITS4 using the DNA polymerase MyFi as previously described, and products were sent to Macrogen (Seoul, Korea) for purification and sequencing. PacBio and Sanger sequences were manually compared in Mesquite v 3.51 (Maddison and Maddison 2017). For samples for which multiple ITS barcodes were recovered for the target species, the sequence versions were compared using Mesquite and the BLASTN suite-2 sequences was used to estimate percentage identity between selected sequence pairs. Sequences with the highest coverage were selected as primary barcodes. Secondary barcodes were only taken into consideration when their coverage was above 30 and when they were less than 98% similar from

the primary barcode. Raw PacBio files and FASTQ files of primary and secondary barcodes obtained were deposited in the Sequence Read Archive on NCBI (BioProject ID PRJNA541190).

Results

Amplifications

After the two amplification steps, 19 of the 96 samples did not show any visible product. For these samples, a new round of amplifications was performed using the non-diluted DNA extracts instead of the 1/10 dilution. This second amplification round generated products for 13 samples that were previously negative. These additional products were then used to substitute the corresponding negative samples on the original plate. For the five samples for which none of the amplification rounds was positive (sample barcode numbers 41, 45, 73, 84, 92, 93), the initial sample were not substituted and were pooled together with the positive samples for final submission.

Sequencing

A first SMRT cell was run with a 0.02 nM sample, but resulting ZMW (Zero-Mode Waveguide) productivity was poor (P1=12%). After filtering, only 17,582 polymerase reads were recovered for this first run, which corresponded to 247,626 subreads (Table 1). The quality of the resulting subreads was, however, acceptable (mean of 14.68 passes). A second SMRT cell was then used with a 0.05 mM concentration of the same sample. The second run showed good ZMW productivity (P1=55%). After filtering, 82,501 polymerase reads were recovered (Table 1). The number of post-filter subreads obtained from this run was 1,095,489 and their quality was comparable to the first run (mean of 13.79 passes). Subreads from both SMRT cell runs were combined and analysed with the Long Amplicon Analysis protocol in the SMRT portal. After demultiplexing, the number of pre-filter subreads per barcode ranged from 91–115,311 (Suppl. material 1: Table S1). Barcodes with low subread numbers (<200) mostly corresponded to samples with negative (41, 45, 73, 93) or weak (53) amplification. The number of CCS obtained for each barcode ranged from

Table 1. Productivity and sequence output from two SMRT runs with different loading concentration.

Loading concentration	ZMW productivity			Post-filter polymerase reads			Subreads		
	P0	P1	P2	number	mean length	N50	number	mean length	number of passes
0.02 nM	86%	12%	2%	17,582	15,234 bp	27,799 bp	247,626	1038 bp	14.68
0.05 nM	29%	55%	16%	82,501	15,321 bp	21,997 bp	1,095,489	1111 bp	13.79

1–948. No chimeric sequences were detected in any of the barcoded samples, but a few low-quality sequences (noise) were found. After clustering and phasing, 0–32 final consensus sequences were obtained per barcode, with a length of 619–2,322 bp, a coverage of 8–500 (the default value of the maximum number of subreads to cluster is 500) and a predicted accuracy of 0.9525–0.9999. The predicted accuracy of the most represented sequence per barcode was above 0.9999 for the majority of the samples (93%).

Sequence identity

Sequences were recovered for 89 of the 96 samples (see Suppl. material 1: Table S1). For each positive sample, 1–32 sequences were obtained using LAA. A BLASTN search was used to identify the sequence of the target species (the lichenised fungi of interest) among the co-amplicons. Eighty-five samples had at least one blast hit that matched the target species. The search revealed that, despite no detection by the PacBio pipeline LAA, a few chimeric sequences (about 3% of the total number of obtained sequences) were obtained, seemingly consisting of two or three full and/or partial sequences, often of the same species, occurring in tandem. Samples for which either no sequence (e.g., samples 41 and 45) or no target species sequences (e.g., samples 92 and 95) principally derived from amplification reactions with no or weak PCR products. Some reactions with weak products did generate a sequence for the target species, although generally with a coverage lower than 100 (e.g., samples 29, 94 and 96).

For 46 samples, a single sequence was obtained for the target species among all the co-amplified fungal taxa. For the other samples, the clustering and phasing process resulted in 2–9 sequences that could be attributed to the target species using the BLASTN match. The similarity between these different sequence versions was investigated in an alignment. After reverse complement and barcode trimming, these sequence versions were in fact identical for an additional ten samples, bringing the total of successful target single-sequences to 56. A few sequence versions also differed with indels in long single nucleotide repeats (e.g., samples 27, 27, 28, 49, 64, 68, 69) and were the likely result of poor quality for low coverage consensus sequences. For the remaining samples, differences between sequence versions were more random (indels and single nucleotide mutations) and more likely due to genuine sequence heterogeneity within a sample, either due to biological reasons (concerted evolution or mixed individual) or carry-over contaminations. The percentage identity between the most divergent target sequence versions of each sample ranged between 94.78 and 99.98% (Suppl. material 1: Table S1).

In addition to the target species, sequences were also obtained for other fungi for 81 of the 96 samples. The great majority of these co-amplifying fungal sequences (479 out of 506 sequences) belonged to the ascomycetes, with Capnodiales and Chaetothyriales being the most commonly represented orders. Several sequences (23) belonged to the

basidiomycetes, two to the chytridiomycetes and one to the mucoromycetes. Finally, one sequence matched *Pirula salina*, a Xanthophyceae algae from the order Tribonematales.

Validation

Sanger sequences were obtained for 12 samples. For five of these (3, 6, 8, 61 and 76), the Sanger sequence was entirely identical to the PacBio sequence. For five additional samples (5, 7, 58, 60 and 74), the Sanger sequence was identical to the PacBio sequence, but contained ambiguous bases or was slightly shorter due to missing bases at the 5' or 3' ends. For two samples (4 and 57), the Sanger sequences comprised one nucleotide difference compared to the PacBio sequences in addition to several ambiguous bases. These differences were located at the start or the end of the sequence and were likely due to poor quality chromatograms in the 5' and 3' regions of the Sanger sequences.

Discussion

Collections of well-curated lichen specimens are important resources for species identification, whether done traditionally using morpho-anatomical together with chemical characters, or using DNA barcodes. For the latter, access to databases with high quality sequence data, detailed voucher information and reliable taxon names is critical. Although current reference sequence datasets exist, they are still incomplete. Thus, the molecular barcoding of lichen specimens for a broad range of species remains a priority (Yahr et al. 2016). In this study, a SMRT sequencing metabarcoding approach generated ITS sequences for 85 of the 96 lichen specimens investigated. This study shows that for lichen herbarium specimens up to 25 years old, full length and high-quality ITS sequences can be recovered for the target lichenised fungus using PacBio amplicon sequencing. Furthermore, sequences were also obtained for co-amplifying fungi, shedding light on the diversity of endolichenic fungi in the samples.

The protocol used in this study enabled the high-throughput processing of samples and the bioinformatic pipeline permitted the recovery of high-quality sequences with minimum time and effort, because both the primary and secondary analyses could be carried out by the sequencing provider. For mixed lichen DNA extracts, SMRT sequencing is a cheaper option than cloning and Sanger sequencing when high number of samples need to be processed. The cost of producing a barcode Sanger sequence is generally around AUD$15/sample (including DNA extraction, amplification and sequencing costs), but when cloning is required, this cost can increase to AUD$100/sample. With the PacBio amplicon protocol used in this study and the PacBio RSII platform, the cost per sample was estimated at AUD$37. This cost could be further reduced by co-amplifying more markers and/or multiplexing more samples. With the new PacBio platform Sequel, multiplexing 384 samples for one marker would bring

the cost down to AUD$12/sample. Moreover, with the recent increased output per SMRT cell (from 1 gigabase for RSII to 20 gigabases for the new PacBio platform Sequel), the cost per sample could fall even further with higher multiplexing levels.

The use of PacBio amplicon sequencing for high-throughput barcoding of lichen specimens is therefore very promising. Some challenges remain, both for the laboratory and the bioinformatic aspects of this method. Some considerations and suggestions for improvements based on this study are outlined below, especially with regard to the preparation of the amplicon library and the generation of ITS barcode sequences using the LAA pipeline.

Technical considerations for high-throughput DNA extraction and amplicon library preparation

A relatively high risk of carry-over contaminations is generally associated with two-step PCR amplicon protocols (Seitz et al. 2015), including the PacBio Barcoded Universal Primers protocol used in this study. In a preliminary test of this protocol (data not shown), output sequences suggested a significant number of potential carry-over contaminations. In light of these results, particular care was taken at various stages of the final run when handling the genomic DNA and PCR samples. More specifically, DNA extracts were stored in single tubes as opposed to a 96-well plate, and amplifications were carried out in strip tubes with individual caps instead of 96-well plates. In the final SMRT run, carry-over contaminations may have also occurred, but at a much lower level than the preliminary test run. Our current protocol could therefore be improved further. More specifically, the use of a PCR hood or post- and pre-PCR separation, as previously recommended (Urban et al. 2000; Jones and Kustka 2017), would be advisable for this two PCR step protocol. During DNA extraction, the use of 8-strip cluster tubes for the grinding of material should be replaced by single tubes, as opening the lids of these tubes causes a potential risk for sample cross-contamination. Minimising the number of pipetting steps and the opening of tubes after amplification is also important. In our protocol, a PCR clean-up step was required after the first PCR due to strong PCR dimer bands. Optimisation of primer concentrations in order to reduce primer dimer formation would eliminate this step. Finally, the inclusion of several replicates of the same sample within a SMRT run would facilitate the assessment of replicability between samples and potential carry-over contaminations.

The choice of polymerase has been shown to be important when using HT sequencing for rare allele detection (Hestand et al. 2016; Potatov and Ong 2017) as well as for community analysis (Oliver et al. 2015), because it impacts directly on the rate of PCR-induced errors. Polymerases with proofreading activities are usually recommended for library preparation protocols, including for SMRT amplicon sequencing. High-fidelity polymerases are, however, more expensive and, in our experience, are more difficult to optimise. A possible incompatibility between a high-fidelity polymerase and SMRT sequencing has even been suggested previously (Schlaeppi et al. 2016).

As argued by Hebert et al. (2018), for the molecular barcoding of reference specimens, the dominant sequence will match the source sequence even when a relatively high number of other sequences have PCR errors. In our preliminary test of this protocol, only a small number of PCR products could be recovered using the high-fidelity Phusion Hot Start II (Thermo Scientific). As the focus of our study was to obtain a barcode sequence from the target species, and not to characterise fungal communities using sequences, the polymerase MyFi was used to generate amplicons. It is marketed as being efficient with challenging templates and inhibitor-rich samples and has so far worked well for lichen specimens. Although not as accurate as a high-fidelity polymerase, MyFi comprises a proofreading component that allows a 3.5× higher fidelity than Taq DNA polymerase.

In addition to polymerase misincorporation, chimera formation is another common PCR-induced source of error. Low starting template concentration and low amplification cycle numbers tend to reduce the rate of chimera formation (Lahr and Katz 2009). Here, 25 cycles were used for both amplifications, and no chimeras were detected using the LLA pipeline. However, after manual inspection of the sequences obtained, several chimeric sequences (about 3%) were detected. Upon closer inspection, these chimeras were found to be formed of concatenated sequences of the full-length amplicon, together with primer sequences and barcodes. These concatemers have previously been reported in SMRT amplicon sequencing studies (e.g., Jones and Kustka 2017), and they are thought to be generated not during the PCR steps, but during the SMRTbell adaptor ligation step of the PacBio library preparation (Fichot and Norman 2013). They can readily be discarded by screening sequences of uncharacteristically large size (Jones and Kustka 2017).

Metagenomics and metabarcoding studies using two-step PCR protocols often include triplicate samples for amplification, which are then pooled and used as a template for the second PCR (e.g., Schlaeppi et al. 2016; Schloss et al. 2016). In community analysis, pooling triplicates is thought to be useful in decreasing PCR bias introduced by inherent differences or stochastic fluctuations in amplification efficiencies (Kennedy et al. 2014), although no study has confirmed that it influences the results of community analyses significantly (Kennedy et al. 2014; Smith and Peay 2014). Using replicates can also reduce the impact of PCR failure due to pipetting-related issues, but it is more expensive and time-consuming. Here, a 'cherry-picking' approach was used where negative samples were redone using a different genomic DNA concentration. Although additional positive samples were recovered with this approach, it may still be possible that PCR product could be recovered from the remaining negative samples using replicates and a broader range of genomic DNA concentrations. Due to large amounts and high diversity of secondary compounds in lichen specimens and their generally low genomic DNA yields, the window of concentration at which amplification is successful is often very narrow, and PCR results are sometimes difficult to reproduce. Using several dilutions of each genomic DNA extract and pooling their products after the first PCR could therefore help in maximising the number of positive samples.

Long Amplicon Analysis and the recovery of high-quality sequences for the target lichenised fungal species

The Long Amplicon Analysis pipeline has been developed by PacBio to phase and detect alleles in diploid organisms, and enables differentiation between underlying sequences that are 99.9% similar, such as haplotypes and pseudogenes (Bowman et al. 2014; Lleras et al. 2014). In this study, as an alternative to cloning, the LAA pipeline was used to recover high-quality barcode sequences from mixed DNA samples of target and non-target fungal species occurring in lichen specimens. The LAA pipeline is accessible on the SMRT portal and can be run by the sequencing service provider together with the primary analysis. The output data generated by LAA included fasta and fastq files with all sequence variants, as well as accuracy and coverage values for each of these sequences. Sequences with the highest coverage generally corresponded to the target species and for more than half of the samples, one single sequence was recovered for the target species. For the other samples, 2–8 sequence versions were recovered per target species. Most often the ITS variation within sample was caused by low quality consensus sequences, which could then be identified due to their low coverage. Occasionally, however, several versions of the ITS barcode with high coverage were recovered within a sample. The percentage identity between these versions mostly ranged between 95–99%. The presence of multiple versions is likely the result of sample heterogeneity (more than one target individual per sample), although carry-over contamination cannot be excluded. Additionally, intragenomic heterogeneity (several ITS copies per nucleus) and intramycelial heterogeneity (several nuclei per mycelium), as discussed in detail in Mark et al. (2016), are two other possible biological reasons for sequence variation. Other ITS sequences obtained corresponded to co-amplifying micro-organisms, predominantly fungi. The most common co-amplified fungi were from the orders Chaetothyriales (Eurotiomycetes) and Capnodiales (Dothideomycetes). The identity of these co-amplified fungi suggested two possible origins for these sequences: either the fungi occur naturally within the lichen thallus (endolichenic fungi), or lichenised or non-lichenised fungal species occur adjacent to the target species on the same substrate, and were accidentally co-sampled during the preparation of material for DNA extraction.

Most target sequences had high accuracy and coverage values, and a subset of these was validated with Sanger sequences obtained from the same DNA extracts. Initially, the size variation among amplicons recovered was more than the 10% recommended by PacBio and was of potential concern. An excess of sequences for the shorter amplicons relative to the longer ones could have been problematic, because lichenised fungi often have long ITS regions due to the presence of introns. However, at this low level of multiplexing (one marker for 96 samples), the difference in amplicon size did not prevent the generation of high-quality sequences for the target species. Moreover, at a higher multiplexing level (5 genes for 96 samples), the sequencing bias due to amplicon size variation did not seem to influence the results (Chen et al. 2015).

A recent study identified some problems with the LAA pipeline, including the formation of a few incorrect or truncated sequences even at relatively high read depths

(Francis et al. 2018). As a result, a new pipeline (C3S-LAA) was developed by these authors which differs from LAA by comparing similarity based on CCSs as opposed to uncorrected reads before the start of the clustering phase. Their new approach, which was used for the SMRT sequencing of long amplicons (4000–8000 bp), successfully eliminated these incorrect and truncated sequences (Francis et al. 2018). We have not observed these problems with our data. However, LAA did not detect chimeras that were formed by concatemers of amplicons with primers and barcodes, sometimes with the second sequence being the reverse complement of the first ("siamaeras", Hackl et al. 2014). However, these concatemers are easily detected with a BLAST comparison and filtered out because of their large size. In addition, some reverse complement sequences (or sequences with slightly truncated barcodes) were not recognised as being identical to other sequences by LAA and were therefore attributed to different clusters. This did not prevent LAA from recovering high-quality sequences for most target species, but it did add some time and effort in verifying whether or not they corresponded to true variants.

Conclusion

PacBio amplicon sequencing is a promising approach for the metabarcoding of lichen specimens, and can be applied to the generation of reference sequences and the characterisation of lichen-associated fungal communities. Although restricted to specimens for which the genomic DNA is not overly degraded, this approach succeeded in generating full-length and high-quality ITS barcodes for specimens up to 25 years old. By scaling up the multiplexing level, this approach could significantly reduce the cost of barcode/sample and compete with Sanger sequencing as well as other NGS approaches.

Acknowledgment

The authors would like to thank Judith Curnow (Australian National Botanic Gardens, Canberra) for her help with specimen databasing and curation, Michel Bertrand and Jean-Yves Monnat for providing lichen specimens, and Lan Li (Australian National Herbarium, Canberra) for her help with the molecular work. We also thank Tonia Russell (Ramaciotti Centre for Genomics, UNSW Sydney, Sydney) and Pacific Biosciences for their advice with sample preparation and data analysis.

References

Anslan S, Bahram M, Hiiesalu I, Tedersoo L (2017) PipeCraft: flexible open-source toolkit for bioinformatics analysis of custom high-throughput amplicon sequencing data. Molecular Ecology Resources 17: e234–e240. https://doi.org/10.1111/1755-0998.12692

Anslan S, Nilsson RH, Wurzbacher C, Baldrian P, Tedersoo L, Bahram M (2018) Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. MycoKeys 39: 29–40. https://doi.org/10.3897/mycokeys.39.28109

Blaalid R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kauserud H (2013) ITS1 versus ITS2 as DNA metabarcodes for fungi. Molecular Ecology Resources 13: 218–224. https://doi.org/10.1111/1755-0998.12065

Bowman BN, Marks P, Hepler NL, Eng K, Harting J, Shiina T, Suzuki S, Ranade S (2014) Long amplicon analysis: highly accurate, full-length, phased, allele-resolved gene sequences from multiplexed SMRT sequencing data. Pacific Biosciences (poster available at https://www.pacb.com/proceedings/long-amplicon-analysis-highly-accurate-full-length-phased-allele-resolved-gene-sequences-from-multiplexed-smrt-sequencing-data).

Chen Y, Frazzitta AE, Litvintseva AP, Fang C, Mitchell TG, Springer DJ, Ding Y, Yuan G, Perfect JR (2015) Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequence typing. Fungal Genetics and Biology 75: 64–71. https://doi.org/10.1016/j.fgb.2015.01.005

Cline LC, Zak DR (2015) Initial colonization, community assembly and ecosystem function: fungal colonist traits and litter biochemistry mediate decay rate. Molecular Ecology 24: 5045–5058. https://doi.org/10.1111/mec.13361

Crous PW, Giraldo A, Hawksworth DL, Robert V, Kirk PM, Guarro J, Robbertse B, Schoch CL, Damm U, Trakunyingcharoen T, Groenwald JZ (2014) The Genera of Fungi: fixing the application of type species of generic names. IMA Fungus 5: 141–160. https://doi.org/10.5598/imafungus.2014.05.01.14

D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N (2016) A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics 17: 55. https://doi.org/10.1186/s12864-015-2194-9

Divakar PK, Leavitt SD, Molina MC, Del-Prado R, Lumbsch HT, Crespo A (2016) A DNA barcoding approach for identification of hidden diversity in Parmeliaceae (Ascomycota): *Parmelia* sensu stricto as a case study. Botanical Journal of the Linnean Society 180: 21–29. https://doi.org/10.1111/boj.12358

Fichot EB, Norman RS (2013) Microbial phylogenetic profiling with the Pacific Biosciences sequencing platform. Microbiome 1: 10. https://doi.org/10.1186/2049-2618-1-10

Francis F, Dumas MD, Davis SB, Wisser RJ (2018) Clustering of circular consensus sequences: accurate error correction and assembly of single molecule real-time reads from multiplexed amplicon libraries. BMC Bioinformatics 19: 302. https://doi.org/10.1186/s12859-018-2293-0

Gardes M, Bruns TD (1993) ITS primers with enhanced specificity for Basidiomycetes: application to the identification of mycorrhizae and rusts. Molecular Ecology 2: 113–118. https://doi.org/10.1111/j.1365-294X.1993.tb00005.x

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics 17: 333–351. https://doi.org/10.1038/nrg.2016.49

Hackl T, Hedrich R, Schultz J, Foerster F (2014) Proovread: large-scale high accuracy PacBio correction through iterative short read consensus. Bioinformatics 30: 3004–3011. https://doi.org/10.1093/bioinformatics/btu392

Hebert PDN, Braukmann TWA, Prosser SW, Ratnasingham S, deWaard JR, Ivanova NV, Janzen DH, Hallwachs W, Naik S, Sones JE, Zakharov EV (2018) A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics 19: 219. https://doi.org/10.1186/s12864-018-4611-3

Heeger F, Bourne EC, Baschien C, Yurkov A, Bunk B, Spröer C, Overmann J, Mazzoni CJ, Monaghan MT (2018) Long-read DNA metabarcoding of ribosomal RNA in the analysis of fungi from aquatic environments. Molecular Ecology Resources 18: 1500–1514. https://doi.org/10.1111/1755-0998.12937

Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR (2016) Polymerase specific error rates and profiles identified by single molecule sequencing. Mutation Research 784/785: 39–45. https://doi.org/10.1016/j.mrfmmm.2016.01.003

Hodkinson B, Lendemer J (2013) Next-generation sequencing reveals sterile crustose lichen phylogeny. Mycosphere 4: 1028–1039. https://doi.org/10.5943/mycosphere/4/6/1

Hofstetter V, Miadlikowska J, Kauff F, Lutzoni F (2007) Phylogenetic comparison of protein-coding versus ribosomal RNA-coding sequence data: A case study of the Lecanoromycetes (Ascomycota). Molecular Phylogenetics and Evolution 44: 412–426. https://doi.org/10.1016/j.ympev.2006.10.016

Irinyi L, Serena C, Garcia-Hermoso D, Arabatzis M, Desnos-Ollivier M, Vu D, Cardinali G, Arthur I, Normand A-C, Giraldo A, da Cunha KC, Sandoval-Denis M, Hendrickx M, Nishikaku AS, de Azevedo Melo AS, Merseguel KB, Khan A, Parente-Rocha JA, Sampaio P, da Silva Briones MR, e Ferreira RC, Muniz M de M, Castanon-Olivares LR, Estrada-Barcenas D, Cassagne C, Mary C, Duan SY, Kong F, Sun AY, Zeng X, Zhao Z, Gantois N, Botterel F, Robbertse B, Schoch CL, Gams W, Ellis D, Halliday C, Chen S, Sorrell TC, Piarroux R, Colombo AL, Pais C, de Hoog S, Zancopé-Oliveira RM, Taylor ML, Toriello C, de Almeida Soares CM, Delhaes L, Stubbe D, Dromer F, Ranque S, Guarro J, Cano-Lira JF, Robert V, Velegraki A, Meyer W (2015) International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database – the quality controlled standard tool for routine identification of human and animal pathogenic fungi. Medical Mycology 53: 313–337. https://doi.org/10.1093/mmy/myv008

James TY, Marino JA, Perfecto I, Vandermeer J (2016) Identification of putative coffee rust mycoparasites via single-molecule DNA sequencing of infected pustules. Applied Environmental Microbiology 82: 631–639. https://doi.org/10.1128/AEM.02639-15

Jones BM, Kustka AB (2017) A quantitative SMRT cell sequencing method for ribosomal amplicons. J. Microbiol. Meth. 135: 77–84. https://doi.org/10.1016/j.mimet.2017.01.017

Kelly LJ, Hollingsworth PM, Coppins BJ, Ellis CJ, Harrold P, Tosh J, Yahr R (2011) DNA barcoding of lichenized fungi demonstrates high identification success in a floristic context. New Phytologist 191: 288–300. https://doi.org/10.1111/j.1469-8137.2011.03677.x

Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD (2014) Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. Applied Environmental Microbiology 80: 5717–5722. https://doi.org/10.1128/AEM.01451-14

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AF, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM, Douglas B, Drenkhan T, Eberhardt U, Dueñas M, Grebenc T, Griffith GW, Hartmann M, Kirk PM, Kohout P, Larsson E, Lindahl BD, Lücking R, Martín MP, Matheny PB, Nguyen NH, Niskanen T, Oja J, Peay KG, Peintner U, Peterson M, Põldmaa K, Saag L, Saar I, Schüßler A, Scott JA, Senés C, Smith ME, Suija A, Taylor DL, Telleria MT, Weiss M, Larsson KH (2013) Towards a unified paradigm for sequence-based identification of fungi. Molecular Ecology 22: 5271–5277. https://doi.org/10.1111/mec.12481

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30: 693–700. https://doi.org/10.1038/nbt.2280

Lahr DJG, Katz LA (2009) Reducing the impact of PCR-mediated recombination in molecular evolution and environmental studies using a new-generation high-fidelity DNA polymerase. Biotechniques 47: 857–866. https://doi.org/10.2144/000113219

Leavitt SD, Esslinger TL, Hansen ES, Divakar PK, Crespo A, Loomis BF, Lumbsch HT (2014) DNA barcoding of brown *Parmeliae* (Parmeliaceae) species: A molecular approach for accurate specimen identification, emphasizing species in Greenland. Organisms Diversity and Evolution 14: 11–20. https://doi.org/10.1007/s13127-013-0147-1

Lleras RA, Bowman B, Tseng E, Wang S, Harting J, Baybayn P, Ranade S, Chin J, Eng K, Marks P (2014) A Novel Analytical Pipeline for de novo Haplotype Phasing and Amplicon Analysis using SMRT™ Sequencing Technology. Pacific Biosciences (poster available at https://www.pacb.com/proceedings/long-amplicon-analysis-highly-accurate-full-length-phased-allele-resolved-gene-sequences-from-multiplexed-smrt-sequencing-data/)

Lücking R, Lawrey JD, Gillevet PM, Sikaroodi M, Dal-Forno M, Berger SA (2014) Multiple ITS haplotypes in the genome of the lichenized basidiomycete *Cora inversa* (Hygrophoraceae): Fact or artifact? Journal of Molecular Evolution 78: 148–162. https://doi.org/10.1007/s00239-013-9603-y

Maddison WP, Maddison DR (2017) Mesquite: A Modular System for Evolutionary Analysis. http://www.mesquiteproject.org [Accessed on 30 May 2018]

Mark K, Cornejo C, Keller C, Flück D, Scheidegger C (2016) Barcoding lichen-forming fungi using 454 pyrosequencing is challenged by artifactual and biological sequence variation. Genome 59: 685–704. https://doi.org/10.1139/gen-2015-0189

Mello A, Napoli C, Morin CME, Marceddu G, Bonfante P (2011) ITS-1 versus ITS-2 pyrosequencing: a comparison of fungal populations in truffle grounds. Mycologia 103: 1184–1193. https://doi.org/10.3852/11-027

Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, Abarenkov K (2010) An open source package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecology 3: 284–287. https://doi.org/10.1016/j.funeco.2010.05.002

Nilsson RH, Taylor AFS, Adams RI, Baschien C, Bengtsson-Palme J, Cangren P, Coleine C, Daniel H-M, Glassman SI, Hirooka Y, Irinyi L, Iršėnaitė R, Martin-Sanchez PM, Meyer W, Oh S-Y, Sampaio JP, Seifert KA, Sklenář F, Stubbe D, Suh S-O, Summerbell R, Svantesson

S, Unterseher M, Visagie CM, Weiss M, Woudenberg JHC, Wurzbacher C, Van den Wyngaert S, Yilmaz N, Yurkov A, Kõljalg U, Abarenkov K (2018) Taxonomic annotation of public fungal ITS sequences from the built environment – a report from an April 10–11, 2017 workshop (Aberdeen, UK). MycoKeys 28: 65–82. https://doi.org/10.3897/mycokeys.28.20887

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research 44: 733–745. https://doi.org/10.1093/nar/gkv1189

Oliver AK, Brown SP, Callaham MA Jr, Jumpponen A (2015) Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15 %. Fungal Ecology 15: 86–89. https://doi.org/10.1016/j.funeco.2015.03.003

Orock EA, Leavitt SD, Fonge BA, St Clair LL, Lumbsch HT (2012) DNA-based identification of lichen-forming fungi: can publicly available sequence databases aid in lichen diversity inventories of Mount Cameroon (West Africa)? The Lichenologist 44: 833–839. https://doi.org/10.1017/S0024282912000424

Potatov V, Ong JL (2017) Examining sources of error in PCR by single-molecule sequencing. PLoS ONE 12: e0168774. https://doi.org/10.1371/journal.pone.0169774

Schlaeppi K, Bender SF, Mascher F, Russo G, Patrignani A, Camenzind T, Hempel S, Rillig MC, van der Heijden MGA (2016) High-resolution community profiling of arbuscular mycorrhizal fungi. New Phytologist 212: 780–791. https://doi.org/10.1111/nph.14070

Schloss PD, Jenior ML, Koumpouras CC, Westcott SL, Highlander SK (2016) Sequencing 16S rRNA gene fragments using the PacBio SMRT DNA sequencing system. PeerJ 4: e1869. https://doi.org/10.7717/peerj.1869

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied Environmental Microbiology 75: 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schoch CL, Robbertse B, Robert V, Vu D, Cardinali G, Irinyi L, Meyer W, Nilsson RH, Hughes K, Miller AN, Kirk PM, Abarenkov K, Aime MC, Ariyawansa HA, Bidartondo M, Boekhout T, Buyck B, Cai Q, Chen J, Crespo A, Crous PW, Damm U, De Beer ZW, Dentinger BTM, Divakar PK, Dueñas M, Feau N, Fliegerova K, García MA, Ge Z-W, Griffith GW, Groenewald JZ, Groenewald M, Grube M, Gryzenhout M, Gueidan C, Guo L, Hambleton S, Hamelin R, Hansen K, Hofstetter V, Hong S-B, Houbraken J, Hyde KD, Inderbitzin P, Johnston PR, Karunarathna SC, Kõljalg U, Kovács GM, Kraichak E, Krizsan K, Kurtzman CP, Larsson K-H, Leavitt S, Letcher PM, Liimatainen K, Liu J-K, Lodge

DJ, Luangsa-Ard JJ, Lumbsch HT, Maharachchikumbura SSN, Manamgoda D, Martín MP, Minnis AM, Moncalvo J-M, Mulè G, Nakasone KK, Niskanen T, Olariaga I, Papp T, Petkovits T, Pino-Bodas R, Powell MJ, Raja HA, Redecker D, Sarmiento-Ramirez JM, Seifert KA, Shrestha B, Stenroos S, Stielow B, Suh S-O, Tanaka K, Tedersoo L, Telleria MT, Udayanga D, Untereiner WA, Uribeondo JD, Subbarao KV, Vágvölgyi C, Visagie C, Voigt K, Walker DM, Weir BS, Weiß M, Wijayawardene NN, Wingfield MJ, Xu JP, Yang ZL, Zhang N, Zhuang W-Y, Federhen S (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. Database 2014: 1–21. https://doi.org/10.1093/database/bau061

Schoch CL, Seifert KA, Huhndorf C, Robert V, Spouge JL, Levesque CA, Chen W, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proceedings of the National Academy of Sciences USA 109: 6241–6246. https://doi.org/10.1073/pnas.1117018109

Seitz V, Schaper S, Dröge A, Lenze D, Hummel M, Hennig S (2000) A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. Nucleic Acids Research 43: e135.

Smith DP, Peay KG (2014) Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. PLoS ONE 9: e90234. https://doi.org/10.1371/journal.pone.0090234

Tedersoo L, Nilsson RH (2016) Molecular identification of fungi, in: Martin, F. (Ed.), Molecular mycorrhizal symbiosis. Wiley-Blackwell, London, 301–322. https://doi.org/10.1002/9781118951446.ch17

Tedersoo L, Tooming-Klunderud A, Anslan S (2018) PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. New Phytologist 217: 1370–1385. https://doi.org/10.1111/nph.14776

Urban C, Gruber F, Kundi M, Falkner FG, Dorner F, Hämmerle T (2000) A systematic and quantitative analysis of PCR template contamination. Journal of Forensic Sciences 45: 1307–1311. https://doi.org/10.1520/JFS14885J

Walder F, Schlaeppi K, Wittwer R, Held AY, Vogelgsang S, van der Heijden MGA (2017) Community profiling of *Fusarium* in combination with other plant-associated fungi in different crop species using SMRT sequencing. Frontiers in Plant Science 8: 2019. https://doi.org/10.3389/fpls.2017.02019

White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics, in: Innis MA, Gelfand DH, Sninsky JJ, White TJ (Eds) PCR Protocols, a Guide to Methods and Applications. Academic Press, San Diego, pp. 315–322. https://doi.org/10.1016/B978-0-12-372180-8.50042-1

Xu M, Heidmarsson S, Thorsteinsdottir M, Eiriksson FF, Omarsdottir S, Olafsdottir ES (2017) DNA barcoding and LC-MS metabolite profiling of the lichen-forming genus *Melanelia*: Specimen identification and discrimination focusing on Icelandic taxa. PLoS ONE 12: e0178012. https://doi.org/10.1371/journal.pone.0178012

Yahr R, Schoch C, Dentinger BTM (2016) Scaling up discovery of hidden diversity in fungi: impacts of barcoding approaches. Philosophical Transactions of the Royal Society B 371: 20150336. https://doi.org/10.1098/rstb.2015.0336

Supplementary material I

Table S1

Authors: Cécile Gueidan, John A. Elix, Patrick M. McCarthy, Claude Roux, Max Mallen-Cooper, Gintaras Kantvilas

Data type: measurement

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mycokeys.53.34761.suppl1